Indonesian Journal of Electrical Engineering and Computer Science
Vol. 21, No. 3, March 2021, pp. 1759~1770
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v21.i3.pp1759-1770

Extracting numerical data from unstructured Arabic texts(ENAT)

Abeer K. AL-Mashhadany[1], Dalal N. Hamood[2], Ahmed T. Sadiq Al-Obaidi[3], Waleed K. Al-Mashhadany[4]
[1,2]Department of Computer Science, Al-Nahrain University, Baghdad, Iraq
[3]Department of Computer Science, University of Technology, Baghdad, Iraq
[4]Iraqi Ministry of Education, Baghdad, Iraq

Article Info

Article history:

Received Jun 12, 2020
Revised Dec 9, 2020
Accepted Dec 20, 2020

Keywords:

Arabic linguistic rules
Numerical dictionary
Related words
Text data mining
Unstructured data

ABSTRACT

Unstructured data becomes challenges because in recent years have observed the ability to gather a massive amount of data from annotated documents. This paper interested with Arabic unstructured text analysis. Manipulating unstructured text and converting it into a form understandable by computer is a high-level aim. An important step to achieve this aim is to understand numerical phrases. This paper aims to extract numerical data from Arabic unstructured text in general. This work attempts to recognize numerical characters phrases, analyze them and then convert them into integer values. The inference engine is based on the Arabic linguistic and morphological rules. The applied method encompasses rules of numerical nouns with Arabic morphological rules, in order to achieve high accurate extraction method. Arithmetic operations are applied to convert the numerical phrase into integer value. The proper operation is determined depending on linguistic and morphological rules. It will be shown that applying Arabic linguistic rules together with arithmetic operations succeeded in extracting numerical data from Arabic unstructured text with high accuracy reaches to 100%.

Corresponding Author:

Abeer K. AL-Mashhadany
Department of Computer Science
Al-Nahrain University
Baghdad, Iraq
Email: aabeeeeraa@yahoo.com

1. INTRODUCTION

Structured data is used by computer programs easily, it is a data in organized form; organized as rows and columns. Unstructured data cannot be used easily by computer programs. It is not in the organized form, for examples; chat-rooms, researchers, images, body of an email, videos, medical reports etc [1-3]. Unstructured text may include within it useful information which is known as unstructured data. The studies proved that more than 80% of business information is found as unstructured data. Unstructured text can be found in; electronic documents, social media, and web pages. It is written in natural languages. For that reason; the two techniques -text mining and natural language processing-are applied in parallel to improve knowledge discovery [4-6].

Text mining is a branch from data mining, so it is text data mining, or it may be called as knowledge discovery in text, or intelligent text analysis. Text data mining derives information from text; which is high quality information. It is considered as one of the most complex analysis because of dealing with unstructured text [7-9]. Natural language processing [NLP] is a subfield of artificial intelligence, which is

concerned with human languages. It tries to provide interactions between human languages and computers. So NLP is necessary and very useful to extract unstructured data [10-12].

Text data mining manipulates unstructured text which includes words, phrases and sentences. It changes them into mathematical values then uses traditional data mining techniques to perform analysis process. These techniques are; information retrieval, information extraction, clustering, categorization, and summarization [13-15]. Natural language processing, together with operation identification techniques reduce problems in text mining procedures. For example, two words may have the same spelling but two different meanings. Linguistic rules are important to describe the entire text [16-18]. M. Zubke [19] extracts numeric values from clinical narratives, and then correct the semantic interpretation. It used regular expression to extract numeric values, also it used template with unambiguous formats or keywords. But there are some difficulties because of the complexity of some numeric values, so machine learning was used to simplify and improve the process of extracting.

T. Cai, et al. [20] developed natural language processing tool, which was simple and powerful, and then performed validations. The developed tool was called EXTEND, it stands for; EXTraction of Electronic medical records Numerical Data. EXTEND had ability to extract numerical physiologic data across different types of notes. It did not need the sophisticated linguistic analysis. It was designed depending on rule-based approach. It had ability to extract numerical information for clinical outcome studies. Both M. Zubke and T. Cai succeeded in extracting numerical values from unstructured text, but that text has a specific format that could be controlled, which means that the two approaches were specialized for clinical report texts, that written in English language. Arabic language achieves level four in the list of most used languages. Natural language processing for Arabic language considered as a challenge because of the complexity of this language [21-23]. Current survey could not find any attempt to extract Arabic numeric words from unstructured texts, instead of that; many attempts had been made to recognize Arabic hand written text.

M. Abuzaraida [24] designed online system for Arabic numerals hand written recognition. Its data set includes 100 samples for each digit written by one writer, it collects its data set from 100 writers. It applied matching alignment algorithm to perform recognition. N. Aouadi [25] proposed automatic system, which extract and recognize Arabic words from hand written examples, while A. El-Sawy [26] proposed a system to extract and recognize Arabic characters.Some previous attempts had been made to implement processes in the branch of natural language processing. ADESS [27], KISB [23], and RSGAS [28] manipulate sentences written in Arabic language. All of them ignore numerical data that may exist within the sentence. Now it is the time to implement the deferred task which is the manipulation of numerical data.

Current work (ENAT) is an attempt to extract numerical data from unstructured Arabic text. ENAT differs from the Arabic examples (M. Abuzaraida, N. Aouadi, and El-Sawy); because it manipulates unstructured text, while they manipulated pictures of hand written examples, also it extracts numerical data, while they recognized and extracted Arabic (digits, words, and characters). ENAT differs from English mentioned examples, because it has ability to extract the numerical data from Arabic numerical words, while they extracted the numerical data from digits only, also it manipulates Arabic unstructured text in general, while they manipulated English unstructured text that is specialized for clinical report. ENAT is developed to achieve a method lost at previous works; ADESS, KISB, and RSGAS. ENAT is important to make attempts of analyzing Arabic text more closed to natural language understanding. ENAT develops a method for analyzing Arabic unstructured text and extract numerical data. Its method depends on numerical Arabic linguistic rules and Arabic morphological rules. In order to implement ENAT method, two main requirements must be provided; dictionary and rules. ENAT dictionary saves all Arabic words that refer to numeric value. ENAT rules includes all Arabic linguistic rules that specialized for numerical branch. These two requirements together give ENAT the ability to perform analysis and extract numerical words.

2. THE PROPOSED SYSTEM TO EXTRACT NUMERICAL WORDS FROM ARABIC TEXTS

The proposed system aims to extract numerical data from Arabic texts (ENAT). It receives Arabic paragraph, analysis it, extract all numerical phrases, and then convert those phrases into integer numbers. The proposed system is decomposed into four main components; two of them are databases; dictionary and rules, while the others are process units; data extraction, and calculation, as shown in Figure 1.

2.1. ENAT Dictionary

ENAT dictionary is dedicated and limited. It includes all Arabic stems that refer to numerical values. Collecting stems required studying some Arabic art articles that concerned with numerical names (*اسماء الاعداد*) in the Arabic language, and manipulate all rules related to numerical names. Steps to build ENAT dictionary are; collecting words, studying the collected words, extracting stems manually, and dividing them into groups according to their linguistic rules. Only stem words are stored. Really there are

many and many Arabic words referred to numerical values, but they not found in ENAT dictionary. That is the reason of providing Arabic morphological rules that concerned with numerical names. So other words will be constructed by applying the rules. The construction task is performed by the data extraction unit.

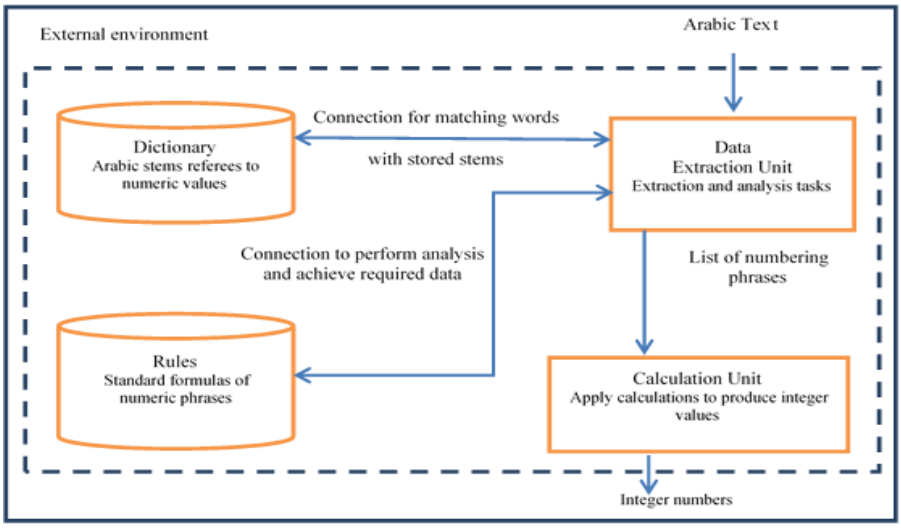

Figure 1. ENAT Architecture

ENAT dictionary can be viewed as a list of compound objects that illustrate stems. The stem compound object stem-co-obj is decomposed into; stem, value, and group name as shown below.Stem-co-obj (stem, value, group), where; ‘stem’ is an Arabic word without prefixes and suffixes that has a numeric value. ‘value’ is the integer number which equals the stem. ‘group’ is the group name to which the stem has belonged.ENAT dictionary completely is shown at Table 1 it grouped words according to their numeric values to facilitate the vision.

Table 1. ENAT dictionary

Value	Stem	group	Value	Stem	Group
1	واحد	Ga1	2	ثنتان	Gb1
	واحدة	Ga2		ثنتين	Gb1
	احد	Ga3		مثنى	Gb1
	احدى	Ga4		اثنان	Gb2
	الأول	Ga5		اثنين	Gb2
	الأولى	Ga5		اثنتان	Gb3
	الحادي	Ga6		اثنتين	Gb3
				الثاني	Gb4
3	ثلاث	Gc1	4	اربع	Gc1
	الثالث	Gc2		الرابع	Gc2
5	خمس	Gc1		رباع	Gc3
	الخامس	Gc2	6	ست	Gc1
7	سبع	Gc1		السادس	Gc2
	السابع	Gc2		ثمان	Gc5
9	تسع	Gc1	8	ثماني	Gc6
	التاسع	Gc2		الثامن	Gc2
10	عشر	Gc4	100	مئة	Ge1
	العاشر	Gc2		مائة	Ge1
20	عشرون	Gd	1000	الف	Ge2
	عشرين	Gd		الاف	Ge3
200	مئتان	Ge2	2000	الفا	Ge3
	مئتين	Ge2		الفان	Ge2
	مائتان	Ge2		الفين	Ge2
	مائتين	Ge2	1000000	مليون	Ge2
	مئتا	Ge4		ملايين	Ge3
	مئتي	Ge4	2000000	مليونا	Ge3
	مائتا	Ge4		مليونان	Ge2
	مائتي	Ge4		مليونين	Ge2
2000000000	مليران	Ge2	1000000000	مليار	Ge2
	مليارين	Ge2		مليارات	Ge3
				مليارا	Ge3

2.2. ENAT Rule

After studying all Arabic linguistic rules that concerned with numerical names; ENAT rule is built to include Arabic numerical morphological formulas (word morphology and phrase morphology). Formulas provide two abilities; constructing words not stored in the dictionary, and testing standardized formulas.ENAT rule includes three parts; part-1 manipulates formulas that have only one word, part-2 manipulates formulas that constructed from two or more words, and part-3 is special for formulas that includes the prefixes 'و' and 'ب'. Before listing ENAT formulas, just notice the following in this work;

a) Group names of ENAT dictionary, are used to write formulas. As an example, Ga1 is an Arabic word belongs to group Ga1, Ga2 is an Arabic word belongs to group Ga2, and so on.
b) The direction of rules is right to left.
c) Symbol + represents the concatenation operation.
d) Symbol bs represents a blank space, in other words; separator between two words.

Part-1 contains formulas from R1 to R7.

R1: Five formulas; include single word that refers to value(1)

Ga1
Ga2
Ga3
Ga4
Ga5

R2: Five formulas; include single word that refers to value(2)

Gb1
Gb2
Gb3
Gb4
Gb4 + ة

R3: Ten formulas; include single word that refers to values(3…10)

Gc1
Gc1+ ة
Gc2
Gc2 + ة
Gc3
Ga4
Gc4 + ة
Gc5
Gc6
Gc6 + ة

R4: Rule of (*الفاظ العقود*), five formulas; include single word that refers to values (20, 30…90) without "ال"

Gd
Gc1 + ون
Gc1+ ين
Gc5 + ون
Gc5 + ين

R5: Rule of (*الفاظ العقود*), one formula; includes single word that refers to values (20, 30…90) with "ال"

ال + R4

R6: Six formulas; include single word that refers to values (100, 200…900, 1000, 2000, 1000000, 2000000, 1000000000, 2000000000) without "ال"

Ge1
Ge2
Ge1 + Gc1
Ge1 + Gc5
Ge3 + Gc1
Ge3 + Gc5

R7: One formula; includes single word that refers to values (100, 200…900, 1000, 2000, 1000000, 2000000, 1000000000, 2000000000) with "ال"

ال + R6

Part-2 contains formulas from R8 to R14.

R8: Four formulas for phrases that refer to value(11)

				Gc4	+	bs	+	Ga3
			ة	+	Gc4	+	bs	+	Ga4
					Gc4	+	bs	+	Ga6
	ة	+	Gc4	+	bs	+	ة	+	Ga6

R9: Four formulas for phrases that refer to value(12)

				Gc4	+	bs	+	ن	-	Gb2
	ة	+	Gc4	+	bs	+	ن	-	Gb3
					Gc4	+	bs	+	Gb4
	ة	+	Gc4	+	bs	+	ة	+	Gb4

R10: Four formulas for phrases that refer to values (13…19) without "ال"

		ة	+	Gc4	+	bs	+	Gc1
		Gc6	+	bs	+	ة	+	Gc1
		ة	+	Gc4	+	bs	+	Gc6
		Gc6	+	bs	+	ة	+	Gc6

R11: One formula for phrases that refer to values (13…19) with "ال"

R10	+	ال

R12: Four formulas for phrases that refer to values (21, 31 …91)

				R4	+	و	+	bs	+	Ga1
				R4	+	و	+	bs	+	Ga4
				R4	+	و	+	bs	+	Ga6
	R5	+	و		+	bs	+	ة	+	Ga6

R13: Four formulas for phrases that refer to values (22, 32 …92)

				R4	+	و	+	bs	+	Gb2
				R4	+	و	+	bs	+	Gb3
				R4	+	و	+	bs	+	Gb4
	R5	+	و		+	bs	+	ة	+	Gb4

R14: Two formulas for phrases that refer to values (23..29, 33..39,.., 93..99)

			R4	+	و	+	bs	+	R3
			R5	+	و	+	bs	+	R3

Part-3 contains the two formulas R15 and R16 to achieve the cases of prefixes "و" and "ب".

R15: Four formulas manipulate the cases of big numbers that multiplied by; thousand, million, or billion. Such formulas may include nested rules that connected by the prefix 'و'

			Ge4	+	bs	+	Ge4	
			Ge4	+	bs	+	R14..R1	
	R14..R1	+	و	+	bs	+	R14..R1	
	R15..R1	+	و	+	bs	+	R15..R1	

R16: One formula for phrases that include the prefix "ب", which is a prefix in Arabic language and may be found as a prefix attached to the first word of a numeric phrase.

R15..R1	+	ب

2.3. ENAT Data Extraction Unit

Data extraction unit receives Arabic paragraphs from an external environment. It analyzes the paragraph, extracts all numerical phrases, constructs a list of them, and passes the list to the calculation unit. Data extraction unit has a direct connection with the dictionary, which is necessary to match the word with numerical stems saved in the dictionary, and then recognizing the numeric word. This connection gives the ability to provide the constructed list with detailed data about each word. All required data to facilitate work of calculation unit will be passed with list of phrases. There is a direct connection between data extraction unit and ENAT rule, which is necessary to perform the analysis task. This connection ensures the verification of numerical phrase. It gives the ability to match the formula of extracted numerical phrases with standard formulas of ENAT rule. That matching is a condition to accept the phrase and then adding it to the list of extracted phrases.

As a preprocessing step, the tokenization process is needed, and then each word will be checked. The job of recognizing numerical word is done by algoritm1. A word may be either numeric word, or not numeric word. There are two possibilities of the numeric word; it may be a word stored in the dictionary, or it may be a stem word attached with prefixes and/or suffixes. The extraction process includes collecting all neighbored numeric words and put them in one phrase, all such phrases will be collected.

Arabic rules of affixes are studied to determine the two lists of prefixes and suffixes; suf-numeric-set = {'ة', 'ون', 'ين', 'مئة', 'مائة'}, and pre-numeric-set = {'ال', 'و', 'وال', 'ب'}. Each one of the two sections (suffix and prefix) may be empty.The output of data extraction unit is a list of numerical phrases,and each phrase is a list of compound objects. For each extracted word, one compound object (word-co-obj) is constructed as shown below.Word-co-obj(suffix, stem, prefix, group, value), where:
'stem', 'suffix', and 'prefix' are compositions of any Arabic word.
'suffix' is an affix belongs to suf-numeric-set.
'stem' is an Arabic stem belongs to ENAT dictionary.
'prefix' is an affix belongs to pre-numeric-set.
'Group' is the ENAT dictionary group, to which the stem belongs.
'Value' is the integer number equal to the stem, taken from dictionary.

```
Algorithm name: Numerical word recognition "algorithm1"
Input: X (Arabic word), D (ENAT dictionary which is list of setm-co-obj), PRE (Pre-numeric-set), and SUF (Suf-numeric-set).
Output: OBJ (word-co-obj).
Processes:
Begin
Step1: F=FALSE        C=1
Step2: DO
       IF X == D[C]. stem THEN
          OBJ = word-co-obj (NULL, D[C]. STEM, NULL, D[C]. GROUP, D[C].VALUE)
          F=TRUE
       ELSE
          IF X== PRE + D[C]. stem + SUF THEN
          OBJ = word-co-obj (PRE, D[C]. STEM, SUF, D[C]. GROUP, D[C].VALUE)
          F=TRUE
       ENDIF
       C=C+1
       UNTIL ((C > length of D) OR (F==TRUE))
Step3: IF F! =TRUE THEN
          OBJ= word-co-obj (NULL, NULL, NULL, NULL, 0)
       ENDIf
End.
```

ENAT developes algorithm2, it extracts the numerical phrase. ENAT develops algorithm3 to implement analysis task. It follows each phrase, word by word, and validates it according to rules belong to the ENAT rule. For each word in standard phrases, a word-co-obj is constructed. In case of a phrase does not match any rule, it will be deleted from the list of phrases.

```
Algorithm name: Numerical phrases extraction "algorithm2"
Input: TEXT (Arabic paragraph).
Output: LNUM (list of lists of word-co-obj).
Processes:
Begin
Step1: L = []
Step2: DO
       X=cut token from TEXT//cut first token of TEXT and put it in X
       OBJ = Call algorithm1(X)
       IF (OBJ.stem! = null) THEN
          L = L U OBJ
       ELSE
            IF (L! = []) THEN
               LNUM = LNUM U L
               L = []
            ENDIF
       ENDIF
UNTIL (TEXT =="") //until end of TEXT
End.

Algorithm name: Numerical phrases validation "algorithm3"
Input: LNUM (list of lists of word-co-objs), and R (list of all ENAT Rules from R1 to R16).
Output: LOBJ (list of lists of word-co-objs).
Processes:
```

```
Begin
Step1: K = 1          F=FALSE
Step2: DO
        NUM= Cut first list from LNUM// cut first number from list of all numbers.
        DO
          IF (NUM Valid to RK) THEN
            F= TRUE
            LOBJ= LOBJ U NUM
          ENDIF
          K=K+1
        UNTIL ((F==TRUE) OR (K>16)
        F=FALSE
        K=1
      UNTIL (LNUM== [])
End.
```

2.4. ENAT Calculation Unit

ENAT calculation unit receives the list of numerical phrases; each phrase is a list of word-co-obj, as passed from the extraction unit. The task of ENAT calculation unit is to convert the list of numerical phrases into integer values. For each numerical phrase; there is only one integer value equal to it.

Really, a numerical phrase consists of related words. The relation is an arithmetic operation, either adding or multiplication. Words that stored in the dictionary have direct values found in the dictionary, as mentioned before; this value is provided to the word-co-obj. Values of Words that recognized by applying morphological rules, require some calculations based on data found in word-co-obj. All data needed to facilitate calculations are saved in word-co-obj.

This section talks about the calculation of one numeric phrase. The same procedure will be applied for all phrases. Simply sub-values will be calculated and saved in a temporary list, and then get the summation of the temporary list. The result of summation represents the final value. To implement the relationship between words and calculate the final value, a structure of work memory is built. It consists of the word-co-obj and set of temporary variables, as shown in Figure 2. Algorithm4 implements task of calculation and result the integer value that equal to the numeric phrase.Algorithmm 4 must be repeated for each one of numeric phrases that are passed from extraction unit.

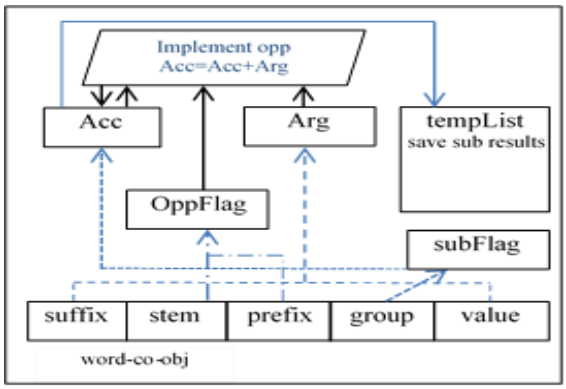

Figure 2. Work memory representation of calculation unit

```
Algorithm name: Integer value calculation “algorithm4”
Input: NUM (list of word-co-obj).
Output: ACC (integer values that equal to the numeric phraes).
Processes:
Begin
Step1: K=1            ACC=0
Step2: OPF = TRUE  SUBF = FALSE
Step3: DO
        DIGIT = Cut first word-co-obj from NUM
        IF ((DIGIT.SUFF =="ون") OR (DIGIT.SUF=="ين")) THEN
          ARG = DIGIT.VALUE * 10
        ELSE
```

```
            IF ((DIGIT.SUFF =="مئة") OR (DIGIT.SUF=="مائة")) THEN
              ARG = DIGIT.VALUE * 100
            ENDIF
          ELSE
            ARG = DIGIT.VALUE
          ENDIF
          IF ((OPF = TRUE) OR (DIGIT.PRE="و") OR (DIGIT.STEM="عشر ") THEN
            ACC = ACC + ARG
          ELSE
            ACC = ACC * ARG
          ENDIF
          IF ((DIGIT.GROUP = "Ge3") OR (NUM= [])) THEN
            SUBF = TRUE
            LSUB[K] = ACC
            K = K+1
            ACC=0
          ENDIF
          IF (SUBF == TRUE) THEN
            SUBF = FALSE
            OPF = TRUE
          ELSE
            OPF = FALSE
          ENDIF
        UNTIL (NUM = [])
    Step4: ACC = Summation of LSUB
    End.
```

3. RESULTS AND DISCUSSIONS

This paper develops dedicated method for numeric-data mining purposes. Researchers aim to achieve successfully a good step in their way to reach natural language understanding. They develope a method lost at their previous works in Arabic text analysis domain. Because all previous works could not manipulate numeric data; in other words, numeric data was neglected at previous works. All other works that were focused during the survey did notdevelpoe such method. ENAT method applied on unstructured text to extract numeric data, it converts numeric words into integer value, while the others manipulate structured text and extract integer data from digits.

ENAT receives Arabic unstructured text. Its job is; searching for numeric words; construct a list of numeric phrases, and then compute and result the integer values that equal to numeric phrases. ENAT applies a good analysis method that has a high degree of accurate,because its analysis depends on Arabic morphological rules,such rules are useful to ensure accuracy, but they considered as complex and needs high level of programming skills. The behavior of ENAT is illustrated in Example 1.

Example 1:

The following paragraph is an Arabic unstructured text. It contains many numeric phrases.

يبلغ عدد السكان في البلد خمسون مليون وتسعمئة الفا واحدى وعشرون نسمة. السكان موزعون في انحاء البلد على النحو الاتي. خمس ملايين وستمئة وسبعون الفا وخمسمئة وخمس وخمسون نسمة في شمال البلد. خمسة عشر ملايين وستون نسمة في وسط البلد. سبعة مليون ومئة الف وثلاث نسمة في شرق البلد. ستة ملايين واربعمئة وخمسون في غرب البلد. سبعة عشر مليون ومئة وثماني وعشرون الفا وتسعمئة وثلاث وخمسون نسمة في جنوب البلد. تبلغ الميزانية المستهلكة للاغراض الصحية عشرة مليارات ومئة واثنان مليون وسبعمئة وستون دينار.

Following is the implementation of ENAT extraction; calling algorithm2. It includes extracting all numerical phrases and construct list of them. It tests each word by implementing algorithm1.

[[خمسون مليونا وتسعمئة الفا واحدى وعشرون], [خمس ملايين وستمئة وسبعون الفا وخمسمئة وخمس]]
[ستة ملايين واربعمئة وخمسون] ,[سبعة مليون ومئة الف وثلاث] ,[خمسة عشر ملايين وستون] ,[وخمسون
[سبعة عشر مليون ومئة وثماني وعشرون الفا وتسعمئة وثلاث وخمسون]
[[عشرة مليارات ومئة واثنان مليون وسبعمئة وستون]

After extraction; analysis must be done. The analysis task is important to ensure the validity of each numerical phrase. The condition to success the ENAT calculation process is the standard form of the input numerical phrase. Following the steps of extraction and analysis process of the first phrase in details.

Step1: Constructing the list of word-co-objs for all numeric words as shown in Table 2. The first field shows the words. The second field shows the suffixes when found. The third field shows the stems. The fourth field shows the prefixes when found. The fifth field shows the group to which the stem is belong. The sixth field shows the integer value of this numeric word.

Table 2.Word-co-obj for each word in phrase خمسون مليونا وتسعمئة الفا واحدى وعشرون

Word	Suff	Stem	Pre	Group	Value
خمسون	ون	خمس	Null	Gc1	5
مليونا	Null	مليونا	Null	Ge3	1000000
وتسعمئة	مئة	تسع	و	Gc1	9
الفا	Null	الفا	Null	Ge3	1000
واحدى	Null	احدى	و	Ga4	1
وعشرون	Null	عشرون	و	Gd	20

Step2: Checking the phrase validity; call algorithm3, if phrase is not in standard form then it must be neglected.

خمسون مليونا وتسعمئة الفا واحدى وعشرون

It is in standard form. It is matching with following formula

واحدى وعشرون				وتسعمئة الف				خمسون مليون
R12	+	bs	+	R15	+	bs	+	R15

For more details

مليون	+	bs	+	خمسون	= خمسون مليون
Ge3	+	bs	+	R4	=R15
الفا	+	bs	+	وتسعمئة	= وتسعمئة الفا
Ge3	+	bs	+	R6	=R15
وعشرون	+	Bs	+	واحدى	=واحدى وعشرون
R4	+	Bs	+	Ga4	R12

Third: The lists of (word-co-obj)s for all phrases will be passed into calculation unit and then have to applyalgorithm4. Table 3 illustrates the simulation of calculation process applied on the first phrase in details, internal variables of ENAT work memory are shown. The output will be the integer number, which is the value of the numerical data extracted from the first numeric phrase.

Table 3.Work memory for calculation of phrase خمسون مليونا وتسعمئة الفا واحدى وعشرون

Step	Acc	Arg	Opp	endSub	subList
Initialize	0	0	+	Off	
خمسون	0	50	+	Off	
	50	50	+	Off	
مليونا	50	1000000	*	Off	
	50000000	1000000	*	On	
	0	1000000	+	Off	50000000
وتسعمئة	0	900	+	Off	
	900	900	+	Off	
الفا	900	1000	*	On	
	900000	1000	+	Off	900000
واحدى	0	1	+	Off	
	1	1	+	Off	
وعشرون	1	20	+	On	
	21	20	+	Off	21
Sum					50900021

Now it is the time to show the ENAT implementation on the other phrases. Table 4 shows implementation of ENAT method on example1 phrases (2..7). For each phrase; firstly the phrase is presented. The analysis task is presented. The word-co-obj for each word in the phrase; is presented. The calculation is presented. And then the integer is resulted.

Table 4. Implementation of analysis and calculation processes on the phrases 2..7

Phrase 2	خمس ملايين وستمئة وسبعون الفا وخمسمئة وخمس وخمسون
Analysis task	خمس / ملايين / وستمئة / وسبعون / الفا / وخمسمئة / وخمس وخمسون R14 + و bs+ + و R6 +Ge3 + R4 + و + R6 +و bs+ Ge3 + bs + R3
word-co-obj	(5، Gc1، ""، خمس،"") (1000000،Ge3،""،ملايين،"") (6،Gc1،و،ست،مئة) (7،Gc1،و،سبع،ون) (1000،Ge3،الفا،""،"") (5،Gc1،و،خمس،مئة) (5،Gc1،و،خمس،"") (5،Gc1،و،خمس،ون)
Calculations	(((10*5)+5)+(100*5) +1000*((10*7)+ (100*6)) + (1000000*5))
The Result	5670555
Phrase 3	خمسة عشر ملايين وستون
Analysis task	خمسة عشر / ملايين / وستون R4 + و +bs + Ge3 + bs +R10
word-co-obj	(5، Gc1، ""،خمس ،ة) (10،Gc4،""،عشر،"") (1000000،Ge3،""،ملايين،"") (6،Gc1،و،ست،ون)
calculation	(10*6)+(1000000* (10+5))
The Result	15000060
Phrase 4	سبعة مليون ومئة الف وثلاث
Analysis task	سبعة / مليون / ومئة / الف / وثلاث R3 +و +Ge3 +و+R6 + Ge3 + bs+ R3
word-co-obj	(7، Gc1، ""، سبع،ة) (1000000،Ge2،""،مليون،"") (100،Ge1،و،مئة،"") (100،Ge2،""،الف،"") (3،Gc1،و،ثلاث،"")
Calculations	3+(1000*100)+(1000000*7)
The Result	7100003
Phrase 5	ستة ملايين واربعمئة وخمسون
Analysis task	ستة / ملايين / واربعمئة / وخمسون R4 +و + R6 + و +Ge3 +bs+R3
word-co-obj	(6، Gc1، ""، ست، ة) (1000000،Ge3،""،ملايين،"") (4،Gc1،و،اربع،مئة) (5،Gc1،و،خمس،ون)
Calculations	((10*5)+(100*4))+(1000000*6))
The Result	6000450
Phrase 6	سبعة عشر مليون ومئة وثماني وعشرون الفا وتسعمئة وثلاث وخمسون
Analysis task	سبعة عشر / مليون / ومئة / وثماني / وعشرون / الفا / وتسعمئة / وثلاث وخمسون R14 +و +و+R6 +Ge3+ bs+R4 +و +و+R3 +R6 +و +Ge3+ bs+R10
word-co-obj	(7، Gc1، ""، سبع، ة) (10،Gc4،""،عشر،"") (1000000،Ge2،""،مليون،"") (100،Ge1،و،مئة،"") (8،Gc6،و،ثماني،"") (20،Gd،و،عشرون،"") (1000،Ge3،""،الفا،"") (9،Gc1،و،تسع،مئة)(3،Gc1،و،ثلاث،"")(5،Gc1،و،خمس،ون)
Calculations	((((10*5)+3)+(100*9))+(1000*(20+8+100))+(1000000*(10+7)))
The Result	17128953
Phrase 7	عشرة مليارات ومئة واثنان مليون وسبعمئة وستون
Analysis task	عشرة / مليارات / ومئة / واثنان / مليون / وسبعمئة / وستون +و R4 +bs+R6 +و +Ge3+ Bs+R2 +و+ bs+R6 +و +bs+R6
word-co-obj	(10، Gc4، ""، عشر، ة) (1000000000،Ge3،""،مليارات،"") (100،Ge1،و،مئة،"") (2،Gb2،و،اثنان،"") (1000000،Ge2،""،مليون،"") (7،Gc1،و،سبع،مئة) (6،Gc1،و،ست،ون)
Calculations	(((10*6)+(100*7))+(1000000*(2+100))+(1000000000*10))
The Result	10102000760

Example 1 includes seven numerical phrases. They covered all possibilities of formulas for Arabic numerical phrases. The implementation of ENAT (extraction, analysis and calculation) method succeeded without any fail. The result of implementation was accurate for all phrases. Example 1 proves that ENAT method provides a perfect analysis procedure with high degree of accurate reaches to 100%. The dependency of analysis procedure on Arabic morphological rules, gives it; the advantage of high accuracy, and the limitation of fail for phrases not written in standard formula. To prove limitation of such procedure, look at example 1, all numeric phrases are written in standard formulas. Any attempt to break the standard formula, it will cause the exception of this phrase from list of numeric phrases, in other words; the lost of its numeric data. If first numeric phrase was written in one of the following not standard foumulas, what will be happen?

خمسون وتسعمئة الفا واحدى وعشرون
خمسون مليونا وتسع واحدى وعشرون
خمسون مليونا تسعمئة الفا احدى عشرون

Such phrase will be exatracted as other phrases, so first step of ENAT method –extraction- will not be affected. Second step is the analysis process. ENAT analysis will marks such phrase as not valid, so it will be deleted from list of extracted phrases. Such phrase will not be found at the last step -calculation-, so its data will be lost.Discussion of example 1 shows that ENAT method is accurate for numeric phrases written in standard formulas, in other side: it can not manipulate not standard formulas. ENAT method presents a good solution to improve the limitation of previous works in text analysis domain. It is good idea to attempt improving ENAT method at future works to be more flexible with popular numeric formulas that used nwadays.

4. CONCLUSION

Text mining is the process of extracting valuable information from unstructured text. Arabic linguistic and morphological rules considered as difficult and need high programming skills, in other side; it ensures high accurate results. Because of the important of results' accuracy; such rules are useful in analyzing unstructured text and extracting numeric data.In this work, all main steps (extraction, verification, and calculation) are dependent on Arabic linguistic and morphological rules, so it achieves results with high accuracy reaches to 100%. Also, minimizing dictionary size and dividing stems into groups; gives advantage point because of minimizing number of required processes. ENAT method is a good step to reach natural language understanding. The future attempts to analyze Arabic unstructured text; will not suffer from difficulty of numeric words, because the proved ENAT method could be used directly. There are two points of limitations for ENAT; firstly, it works only with integer numbers, and secondly, it works only with text written based on standard linguistic rules. It is good idea for future work to improve ENAT method. So, it will be more flexible with popular numeric formulas that used nwadays.

ACKNOWLEDGEMENTS

The authors would like to thank Al-Nahrain University, University of Technology and Iraqi Ministry of Education for their support to conduct the work published in this paper.

REFERENCES

[1] J. Wagh, *et al.*, "Unstructured Data Mining and Its Applications," *International Journal of Current Engineering and Scientific Research (IJCESR),* vol. 3, no. 3, pp. 36-40, 2016.
[2] K. Kanimozhi, *et al.*, "Unstructured Data Analysis-A Survey," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 4, no. 3, pp. 223-225, 2015.
[3] Z. Rybchak, *et al.*, "Analysis of methods and means of text mining", *ECONTECHMOD: an International Quarterly Journal on Economics of Technology and Modelling Processes,* vol. 6, pp.73–78, 2017.
[4] F.S.Gharehchopogh, *et al.*, "Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing," in *5th International Conference on Application of Information and Communication Technologies, 2011. AICT 2011,Baku, IEEE,* 2011, pp. 1-4.
[5] A. Yassi, *et al.*,"Sentimental classification analysis of polarity multi-view textual data using data mining techniques," *InternationalJournal of Electrical and Computer Engineering (IJECE),*vol. 10, no. 5, pp. 5526-5534, 2020.
[6] A. Alhad, *et al.*, "A computational analysis of short sentences based on ensemble similarity model," *International Journal of Electrical and Computer Engineering (IJECE),*vol. 9, pp. 2088-8708, 2019.
[7] N. Kamaruddin, *et al.*, "Jobseeker-Industry Matching System Using Automated Keyword Selection and Visualization Approach," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 13, no. 3, pp. 1124-1129, 2019.
[8] B. Banazir and P. Annes, "Efficient Keyword Search Using Text mining Techniques: a Survey," *International Journal of Engineering and Innovative Technology (IJEIT),* vol. 9001, no. 1, pp. 2277-3754, 2013.
[9] K. Prakash, *et al.*, "Advances in Natural Language Processing – A Survey of Current Research Trends, Development Tools and Industry Applications," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 7, pp. 199-201, 2019.
[10] P. Mayr,*et al.*, "Introduction to the special issue on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)," *International Journal on Digital Libraries,*vol.19, no. 2-3, pp. 107–111, 2018.
[11] V. Venkatesh, "Accelerating Information Retrieval using Natural Language Processing," *International Journal of Computer Science Trends and Technology (IJCST),* vol. 6, no. 3, pp.117-132, 2018.
[12] A. Al-Mashhadany, *et al.*, "Textual Analysis Applications: Subject Review," *Journal of university of Anbar for Pure science,*vol. 12, no. 3, pp. 71-83, 2018.
[13] S. Sangam, *et al.*, "Sentiment classification of social media reviews using an ensemble classifer," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 16, no. 1, pp. 355-363,2019.
[14] H. Hassani, *et al.*, "Text Mining in Big Data Analytics," MDPI. *Big Data and Cognitive Computing,* vol. 4, no. 1, pp. 1-34, 2020.
[15] S. Tandel, *et al.*, "A Survey on Text Mining Techniques," in *5th International Conference on Advanced Computing & Communication Systems, 2019. ICACCS 2019, Coimbatore, India, IEEE,* 2019, pp. 1022-1026.
[16] M. Biniz, *et al.*, "Ontology Matching Using BabelNet Dictionary and Word Sense Disambiguation Algorithms," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 5, no. 1, pp. 196- 205, 2017.
[17] M. Fikri, *et al.*, "A comparative study of sentiment analysis using SVM and SentiWordNet," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 13, no. 3, pp. 902-909, 2019.
[18] G. Coro, *et al.*, "An Open Science System for Text Mining," in *Proceedings of the Sixth Italian Conference on Computational Linguistics, 2019. CliC-it 19,Bari, Italy, CEUR-WS.org,*vol. 2481, no. 23, 2019.
[19] M. Zubke, "Classification based extraction of numeric values from clinic alnarratives," in *Proceedings of the Biomedical NLP Workshop associated with RANLP, Varna, Bulgaria, INCOMA Ltd,* 2017, pp. 24–31.

[20] T. Cai, *et al*., "EXTraction of EMR numerical data: an efficient and generalizable tool to EXTEND clinical research,"*BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 226, 2019.https://doi.org/10.1186/s12911-019-0970-1
[21] A. Jihad, et al, "A framework for sentiment analysis in Arabic text," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1482-1489, 2019.
[22] I. Guellil, *et al.*, "Arabic Natural Language Processing: an Overview," *Journal of King Saud University – Computer and Information Sciences*, 2019. https://doi.org/10.1016/j.jksuci.2019.02.006
[23] A. Al-Mashhadany, "Knowledge Acquisition & Hybrid Inference with a Stem-Based Approach (KISB)," *Journal of Al-Nahrain University - Science*, vol.15, no. 2, pp.169-183, 2012.
[24] M. Abuzaraida, *et al.*, "Online Recognition Approach of Arabic Numerals Using Matching Alignment Algorithm," *International Journal of Data Science and Analysis*, vol. 2, no. 2, pp. 37-41, 2016.
[25] N. Aouadi, "Word Extraction and Recognition in Arabic Handwritten Text," *International Journal of Computing and Information Sciences*, vol. 12, no.1, pp. 17-23, 2016.
[26] A. El-Sawy, *et al.*, "Arabic Handwritten Characters Recogniting Using Convolutional NeuralNetwork," *Wseas Transactions On Computer Research*, vol. 5, pp.11-19, 2017.
[27] A. Al-Mashhadany, *et al*., "Arabic Diagnosing Expert System Shell (ADESS)," in *The Proceedings of 2nd Conference for Information Technology: Applications and Horizons, University of Technology, Baghdad, Iraq*, 2010, pp. 172-193.
[28] A. Al-Mashhadany, *et al*., "Root-Stem Approach in General Analyzer System for Arabic Language (RSGAS)," *Journal of University of Al-Anbar for pure science*, vol. 10, no. 3, pp. 41-49, 2016.

BIOGRAPHIES OF AUTHORS

Abeer Khalid Al-Mashhadany received her BSc in Computer Science in 2002 and MSc in Computer Science (Artificial Intelligence) in 2005 from Al-Nahrain University, Baghdad, Iraq. She is Assistant Prof. in A.I. since 2015. Her research intrests include artificial intelligence, natural language processing, and analysis of arabic and english texts.

Dalal Naeem Hammod received her BSc in Computer Science in 2002 and MSc in Computer Science (Network Security) in 2005 from Al-Nahrain University, Baghdad, Iraq. Her Ph. D from University of Technology (Network Security and Artificial Intelligence) in2018. She is lecturer since 2011. Her research intrests include network security, autonomous security, authentication, modren encryption, multi-agents' system, and artificial intelligence.

Ahmed T. Sadiq received a B.Sc., M.Sc. & Ph. D. degree in Computer Science from the University of Technology, Computer Science Department, Iraq, 1993, 1996 & 2000 respectively. He is Professor in A.I. since 2014. His research interests in artificial intelligence, data security, patterns recognition and data mining.

Waleed Khalid Al-Mashhadany received his BSc in Literature of the Arabic language in 1991 from Al_Mustansiriyah University. He worked as a teacher at Ministry of Education/ Iraq (1991-2019). His research interests were in Arabic langusge. He recently died of cancer.